**鐵達尼號生存預測模型**

|  |  |  |
| --- | --- | --- |
| 曾凰嘉 | 林家妤 | 陳縵欣 |
| 國立中山大學 | 國立中山大學 | 國立中山大學 |
| B063022025 | B063040056 | B063040059 |

|  |  |  |
| --- | --- | --- |
| 陳少洋 | 王譽鈞 |  |
| 國立中山大學 | 國立中山大學 |  |
| B063040061 | B065040034 |  |

# 摘要

　　鐵達尼號沈船事件使得Kaggle網站收集了鐵達尼號每位乘客的相關資料，並且創造了"Titanic:Machine Learning from Disaster"競賽，此競賽主要希望參賽者透過資料分析並且做適當的做資料預處理，以及建立學習模型使機器學習，並能夠預測出鐵達尼號乘客的存亡(標籤)。由於在資料處理的部分有些欄位缺失值，就要依據乘客的其他資料去預測這些缺失值，或是有些欄位的資料過於複雜，因此必須擷取其中與存亡最有關聯的數個特徵，此外，我們分析了那些特徵是否有參考的必要性，若能在資料預處理上做的越完善就有機會得到較好的結果。

**關鍵詞：鐵達尼號生存預測、Kaggle、機器學習。**

## 簡介

人工智慧的技術不斷的進步，透過機器學習可以讓機器用來幫人類判斷許多情境，例如：偵測疾病、alphago、車牌辨識等，因此我們利用機器學習來了解鐵達尼號事件，鐵達尼號的沉船之謎一直以來都引發人們強烈的興趣和熱絡的討論，當時如此巨大的一艘郵輪居然因為撞上了冰山而導致沉船，沉船前，乘客們紛紛搭乘救生艇逃亡，但全部的救生艇也無法容納下船上的所有乘客，加上半夜溫度急遽下降，造成最後只有部分成功倖存，有些人隨著鐵達尼號深沉大海之中，為瞭解當時擁有那些條件的人存活的可能性較大，我們利用鐵達尼號上部分乘客的資料，透過機器學習的方式做訓練並且預測擁有哪些條件的人會生存下來，或許就能夠更了解當時沉船前乘客逃難情形。

## 相關研究

研究[1]提到ticket特徵其實跟乘客的生存有很大的關係，當ticket內有"PC 17755”這個字串出現時，乘客全部活了下來，而開頭是"A” 、”W”、”SOTON"的乘客存活的機率就比較小，大部分都死掉了，這可能是因為ticket前面的編號跟乘客在船上的位置有關係，ticket有"PC 17755”的乘客可能較靠近船的邊緣，因此逃難時就可以較快到達救生艇，此外，該研究提到家庭成員數對於生存率是沒有太大關聯，但我們認為家庭成員數量仍有參考價值，因為這些跟親戚一起上船的人，他們最後幾乎一起死亡，或是一起存活，因此我們認為家庭成員數就有助於預測，而fare特徵在文獻中，則被認為是沒有參考價值，因為它會使最後預測的準確率降低。

## 程式設計方式

* 1. **資料前處理**

PassengerID：本次模型設計我們假設乘客編號與存活無關，因此在做資料訓練時不保留此欄位。

Survived：由於此欄位屬於訓練資料的標籤(label)，因此保存下來用來計算預測準確度，而在做資料訓練時不保留此欄位。

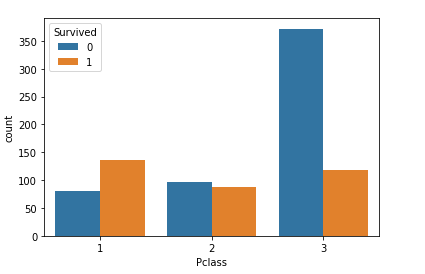
Pclass：依據不同Pclass對survived會有不同比例的影響(圖1)，根據這個比例去值，Pclass=1 的存活率有7成左右，Pclass=2 有5成，Pclass=3 只有3成左右的存活率。

圖1。艙等與存亡的統計

Gender：由於需特徵化，性別也須轉成數字，但是不能填0和1，因為這樣將可能會代表有或沒有的意思，會大幅影響中間過程的計算，所以採用one hot encoding的方式編碼。(圖2)

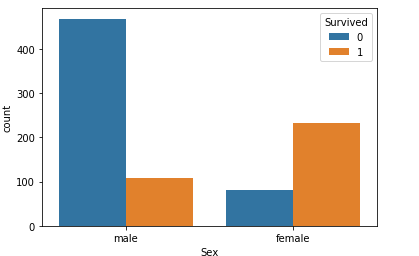


圖2。性別與存亡的統計

Name：我們將把前面的稱謂給留下，後面的名字則捨棄掉之後，發現有很多種稱謂，但是有些稱謂的名稱比較少，導致發生資料不平衡(Imbalanced Data)的情況，因此我們假設：若訓練資料該樣本人數特別少時，將可能會影響準確預測。例如，提供的訓練資料集姓名欄位只有一位Sir.稱謂的資料，但我們大膽假設一個資料並無法代表所有的Sir.。我們就把它分配到稱謂比較多的類別，舉例來說，Dr(資料集出現較少)就分配給Mr(資料集出現較多)，而Lady分配給Mrs，最後只留下稱謂比較多的Master、Miss、Mr、Mrs。

Ticket：Ticket前面的英文字母可能代表房間的位置，後面的數字猜測沒有意義所以省略，以空白作為英文字母和數字的區別，把斜線(/)跟句號(.)的符號也去除掉，如果是全部都數字的話就填上’X’，就留下A5、PC、STONO2等等。

Age：依據每個乘客姓名的稱謂，從訓練集計算出該稱謂其餘乘客的平均年齡，並按照該稱謂計算出的平均年齡去填補其乘客所對應缺失的年齡值。(圖3)

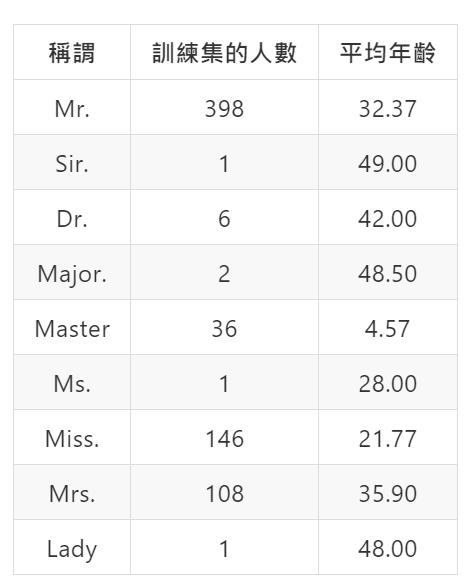


圖3。該稱謂的乘客年齡無缺失之平均年齡與稱謂人數

Cabin：取cabin前面的英文字母，並將後面的數字捨去，只留下乘客位在哪一個Cabin，而缺失值填 ’X’。

Fare：我們假設價格會與Pclass(艙等level)及出發地(Embark)有關，因此保留Pclass和Embarked就好，因此Fare須捨去掉。

Sibsp與Parch：從相關係數可以看出如果Sibsp和Parch的人數越多，其實存活的機率越大，因此我們做特徵合成，將把Sibsp、與Parch兩個欄位合併成新的欄位：家庭成員數(Family)，其數值為Sibsp、與Parch兩個欄位的數值相加。

Embarked : 把出發地的英文字母轉成one hot encoding的形式編碼。

* 1. **演算法**

　　本次模型採用全連接神經網路 (Fully-connect Neural Network, FNN)，分成輸入層 (Input layer)、隱藏層 (Hidden layer) 以及輸出層 (Output layer)，也就是說，若要計算下一層某顆神經元，則須計算出與該層連接的上一層所有神經元的輸入值與其神經元權重相乘之總和，並將結果值作為的該顆神經元的值，在計算過程中程式會反覆調整權重(weight)和偏差值(bias)，其過程稱為反向傳播 (Backpropagation)。

　　隱藏層之激勵函數採用ReLU函數，其意義為當計算結果值小於等於0，其對應值為0，計算結果大於零時，其對應值不變。而輸出層採用softmax激勵函數，讓輸出的值介於0~1之間，使該層2顆神經元的輸出(也就是機率)總和為1。

　　此外，損失函數(Loss Function)採用了Cross Entropy，屬於分類的常用函數，目的是為了獲得最大的資訊增益(information gain)，同時希望能夠盡量符合真實的資料分布，以降低泛化誤差(Generalization error)。

* 1. **問題與討論**

　　一開始我們只做簡單的資料處理，發現準確率很低，於是開始針對每個欄位的資料做分析，例如Pclass這個欄位一開始我們是想用one hot encoding的方式去處理，因為如果如果給程式1、2、3的數值，程式可能會認為這是對survived分別造成1倍、2倍、3倍的關係，但這並不是資料想表達的意思，用了one hot encoding後發現其實準確率沒差很多，於是我們就先印出Pclass跟survived之間的關係，發現Pclass=1和2的生存機率比較大，Pclass=3的存活機率較小，於是就依據Pclass數值和survived之間的關係給值，發現可以得到較好的結果。此外，我們從文獻結果[1]得知Ticket的開頭為"A"、'W"、"SOTON"有較大的機率死亡，因此將這些人的Ticket改成較小的數字0，和Ticket是"PC 17755”的乘客全部存活，因此將Ticket值改成較大的數字1，其他的Ticket值就取一個較中間值為0.5，但經過處理後準確率其實仍然相差不大，因此我們認為可能是Ticket取中間值得資料太多了，可能要再分析更多資訊，把這些中間值再改成更準確的數字，或許準確率就可以再增加一點。

## 實作與實驗結果

* 1. **實作**

　　本次模型使用Python3程式語言實作，其包含tensorflow、pandas、numpy、keras套件。資料前處理階段，我們將Age、Cabin的欄位補上缺失值，其中Age是透過Name的稱謂來判別，在標題3.1提過處理細節，而Cabin直接將缺失值補上X。在訓練階段，我們將SibSp、Parch合併成Family(家庭成員數)特徵，並且將Name(只取稱謂)、Ticket、Fare、Pclass(分為1、2、3)、Sex、Age、Embarked(以one hot encoding形式呈現C、Q、S之不同的登船口)保留，其餘特徵捨棄(圖4)，因此共有12個特徵，作為神經網路之輸入層神經元的個數(圖5)。

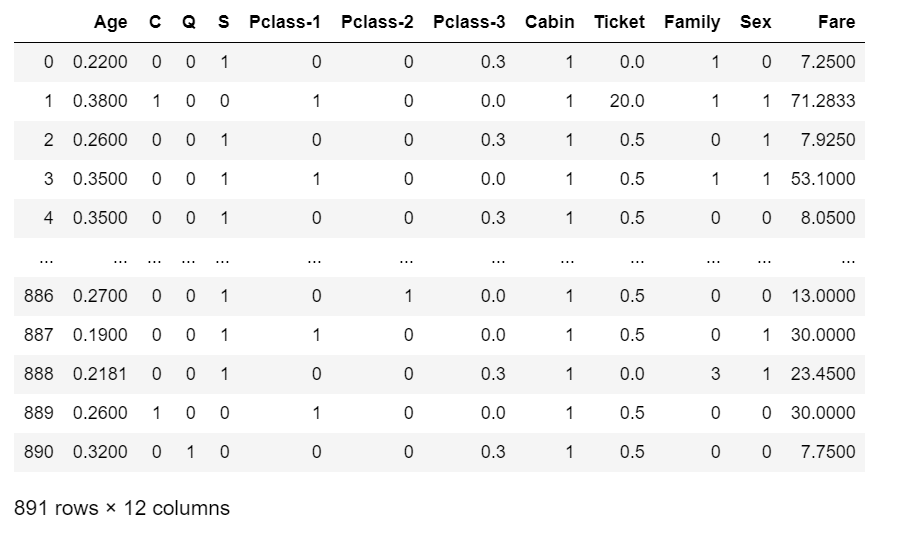


圖4。訓練的資料特徵。

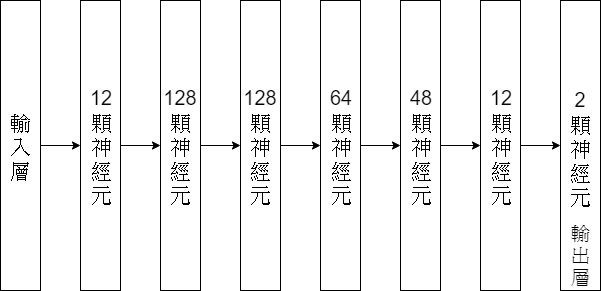


圖5。神經網路架構

* 1. **實驗結果**

1.　本模型採用全連接神經網路，其Kaggle準確率判斷最高為78.468%。(圖6)

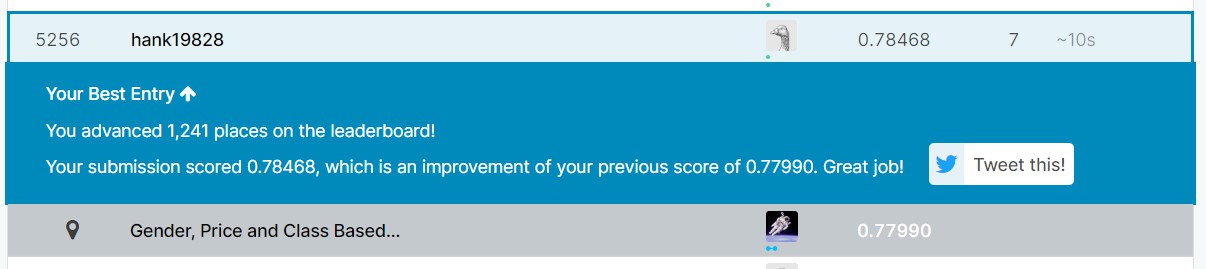


圖6。Kaggle競賽網站排名與準確率

2.　研究[1]提到若將Fare特徵捨棄並且使用高斯樸素貝葉斯分類器(Gaussian naive Bayes)可大幅提高準確度(根據文獻提供為92.5%)，但我們使用神經網路學習，實驗後只有77.511%，反而若保留Fare欄位準確率提高成78.468%。

3.　實驗過程使用兩種激勵函數，結論為使用LeakyReLU函數與ReLU無顯著差別。

4.　若將Sex特徵捨棄掉，則Kaggle準確率下降為54.066%，也就是說性別與存亡有顯著關聯，因此必須保留。

5.　若將Embarked特徵捨棄掉，則Kaggle準確率微降為77.511%，因此登船口與存亡無顯著關聯，因此可保留也可捨棄。

## 結論

　　我們認為，採取不同的學習模型，可能會因為保留及捨棄哪些特徵而影響最終的準確度。此外，根據文獻[1]指出，使用支撐向量機(support　vector　machine, SVM)模型來訓練鐵達尼號乘客資料大多落在77%，而使用隨機森林(Random　Forests)分類準確率落在81%，本模型採用全連接神經網路，其Kaggle準確率最終判斷為78.468%。

## 參考文獻

[1]. Nadine Farag, Ghada Hassan, Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster, May 2018